CompGuessWhat?!
A Multi-task Evaluation Framework for Grounded Language Learning
Suglia, Alessandro; Konstas, Ioannis; Vanzo, Andrea; Bastianelli, Emanuele; Elliott, Desmond; Frank, Stella; Lemon, Oliver

Published in:
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

DOI:
10.18653/v1/2020.acl-main.682

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning

Alessandro Suglia\textsuperscript{1}, Ioannis Konstas\textsuperscript{1}, Andrea Vanzo\textsuperscript{1}, Emanuele Bastianelli\textsuperscript{1}, Desmond Elliott\textsuperscript{2}, Stella Frank\textsuperscript{3}, and Oliver Lemon\textsuperscript{1}
\textsuperscript{1}Heriot-Watt University, Edinburgh, UK
\textsuperscript{2}University of Copenhagen, Copenhagen, Denmark
\textsuperscript{3}University of Edinburgh, Edinburgh, UK
\textsuperscript{1}\{as247,i.konstas,a.vanzo,e.bastianelli,o.lemon\}@hw.ac.uk
\textsuperscript{2}de@di.ku.dk
\textsuperscript{3}stella.frank@ed.ac.uk

Abstract

Approaches to Grounded Language Learning typically focus on a single task-based final performance measure that may not depend on desirable properties of the learned hidden representations, such as their ability to predict salient attributes or to generalise to unseen situations. To remedy this, we present GrOLLA, an evaluation framework for Grounded Language Learning with Attributes with three sub-tasks: 1) Goal-oriented evaluation; 2) Object attribute prediction evaluation; and 3) Zero-shot evaluation. We also propose a new dataset CompGuessWhat?! as an instance of this framework for evaluating the quality of learned neural representations, in particular concerning attribute grounding. To this end, we extend the original GuessWhat?! dataset by including a semantic layer on top of the perceptual one. Specifically, we enrich the VisualGenome scene graphs associated with the GuessWhat?! images with abstract and situated attributes. By using diagnostic classifiers, we show that current models learn representations that are not expressive enough to encode object attributes (average F1 of 44.27). In addition, they do not learn strategies nor representations that are robust enough to perform well when novel scenes or objects are involved in gameplay (zero-shot best accuracy 50.06%).

1 Introduction

Several grounded language learning tasks have been proposed to capture perceptual aspects of language (Shekhar et al., 2017; Hudson and Manning, 2019; Suhr et al., 2019; Agrawal et al., 2018). However, the advances in this field have been primarily driven by the final performance measures and less on the grounding capability of the models. In fact, in some cases, high-performance models exploit dataset biases to achieve high scores on the final task (Zhang et al., 2016; Agrawal et al., 2016). In the literature, several methods have been proposed to analyse what kind of information is captured by neural network representations (Kádár et al., 2017; Belinkov and Glass, 2019). Most of these works examine the hidden state representations learned by models trained on only textual data. However, many aspects of human semantic representations are grounded in perceptual experience (Andrews et al., 2009; Riordan and Jones, 2011). This paper explores the idea that visually grounded representations ought to be a result of systematic composition of grounded representations (Harnad, 1990). For instance, the understanding of the word “microwave” is grounded in perception of objects with specific attributes such as shape, colour, and size – see Figure 1 for an example. Therefore, investigating whether the representations learned by a model exhibit forms of attribute composition is beneficial for assessing model interpretability and generalisation.

In this work, we propose GrOLLA – a multi-task evaluation framework for Grounded Language Learning with Attributes that expands a goal-
Figure 2: CompGuessWhat?!: Detailed description of the attributes of two different objects in the reference scene. Both the objects have a set of abstract attributes (indicated in blue) and a set of situated attributes (indicated in green).

Oriented evaluation – based on the standard final task measure, with two auxiliary tasks: 1) Object attribute prediction (AP), and 2) Zero-shot evaluation (ZS). The attribute prediction task is designed to evaluate the extent to which the model’s latent representations associated with objects are useful for predicting their attributes. The prediction performance on this task can be related to a degree of compositionality of the learned representations. We adopt a behavioural, i.e., task-driven, approach to assessing aspects of compositionality for visually grounded representations, whereby the extent to which a representation is compositional depends on: (a) its ability to predict object attributes, and (b) its ability to generalise to novel contributions of object attributes. To support (b), we design a zero-shot evaluation that measures the extent to which the learned representations can be reused in a task involving objects unseen during training. By optimising for both the final end-goal measure as well as the auxiliary tasks, we aim to drive the design of models that can solve the task more reliably and whose representations are easier to interpret as a result of being a composition of visual attributes.

This paper presents three main contributions: (1) We define GroLLA – a multi-task evaluation framework for grounded language learning that augments the final end-goal measure(s) with auxiliary tasks aimed at assessing the degree of attribute grounding of the model’s representations; (2) We propose an instance of this multi-task evaluation framework, namely CompGuessWhat?!; and (3) We evaluate state-of-the-art models using the CompGuessWhat?! dataset. The evaluation shows that models with high performance in the end-goal task are not able to reliably predict the attributes of given objects and do not generalise to examples with unseen object categories.

CompGuessWhat?! is a benchmark of 65,700 dialogues (see Section 3). It is based on GuessWhat?! (de Vries et al., 2017) dialogues and enhanced by including object attributes coming from resources such as VISA attributes (Silberer and Lapata, 2012), VisualGenome (Krishna et al., 2017) and ImSitu (Yatskar et al., 2016).

2 Evaluation Framework

Our evaluation framework for Grounded Language Learning tasks is based on three different sub-tasks: 1) Goal-oriented evaluation; 2) Object attribute prediction evaluation; 3) Zero-shot evaluation.

Goal-oriented evaluation We evaluate the models according to the multi-modal task that they have to solve, which can generally be categorised as classification or generation. Classification tasks such as Visual Question Answering (Antol et al., 2015) or Visual Natural Language Inference (Suhr et al., 2019) involve predicting the correct label for a given example whose performance is measured in terms of predictive accuracy. In generative tasks, such as Image Captioning (Bernardi et al., 2016), the model has to learn to generate a sequence of labels for a given input data whose performance measure is BLEU (Papineni et al., 2002).

Object attribute prediction evaluation We support the goal-oriented evaluation with the attribute prediction auxiliary task related to assessing the degree of compositionality of the representations learned for a specific task.

With an attribute prediction task, we can assess whether the learned representations capture what we think they should, in terms of object attributes, rather than spurious correlations. The idea of using object attributes as an auxiliary task follows from the Characteristic Feature Hypothesis (Hampton, 1979) according to which every concept category has a set of defining features, which provide a criterion for judging which objects are category members, and which are not. Therefore, the higher the accuracy in the attribute prediction task, the more the representations learned by the model are composed of the set of attributes of the objects.

Zero-shot Evaluation Via the attribute prediction task, we can assess the ability of latent representations to recover some of the attributes associated with their object category. Assuming that the model has learned to represent these attributes, we
We enrich the VisualGenome (Krishna et al., 2017) with several attributes coming from resources such as VISA (Silberer and Lapata, 2012) and ImSitu (Yatskar et al., 2016). Unfortunately, not all the GuessWhat?! images are included in VisualGenome. We were able to reuse 40.79% of the original GuessWhat?! dialogues for a total of 65,700 dialogues (additional information can be found in the related Appendix A.1). By relying on this set of attributes, we define an attribute prediction evaluation to assess the extent to which the learned neural representations can encode the attributes specified during the dialogue. In order to determine the generalisation power of the learned representations and their ability to be transferred, we propose a novel zero-shot learning set of reference games involving target object belonging to an unseen object category. The dataset and the code associated with this paper can be found online1.

Psycholinguistically-motivated attributes We extend the set of attributes for every object category in MSCOCO with psycholinguistically-motivated semantic representations based on the McRae Norms (McRae et al., 2005) developed by Silberer and Lapata (2012). We use only the subset of so-called abstract attributes, and ignore attributes from the original set that can change depending on the reference image (e.g., “shape”, “texture”, etc.). We use the WordNet synset identifier (e.g., “person.n.01”) associated with a given MSCOCO category (e.g., “person”) to automatically associate its corresponding abstract attributes with a specific object instance. However, very often several Visual Genome objects have a synset associated with a class that is a hyponym of the MSCOCO category synset. Therefore, we rely on the Wu-Palmer similarity (Wu and Palmer, 1994) to find the best match between the Visual Genome synset and the MSCOCO category synset (with a similarity threshold of 0.75 chosen by using as reference the distance between the synset of person and woman). The intuition behind this heuristic is that we assume that a hyponym will inherit the abstract attributes of its hypernym.

Affordances & Behaviours We extract the semantic roles associated to specific object categories using the ImSitu dataset (Yatskar et al., 2016), in order to include affordances and behaviours associated with every object category. An object category is associated with a behaviour every time it appears as the agent of a given predicate. For in-

---

1https://compguesswhat.github.io
We consider the guesser accuracy metric (in gameplay mode) from the GuessWhat?! dataset for our goal-oriented evaluation. It measures how many times the guesser model can select the correct target object among the candidate objects, given the dialogue generated so far. Due to the importance of this language game for NLU and NLG model skills, we decide to keep the guesser accuracy as a reference metric to assess the ability of the questioner to play the game. However, unlike the original dataset evaluation, we make sure that the score is evaluated ignoring duplicated dialogues.  

4.2 Attribute Prediction Evaluation

In a sequential guessing game like the one in Figure 1, we regard the representation for the last turn of the dialogue as a composition or aggregation of all the attributes specified so far. Therefore, we can use it to predict with high accuracy the attributes associated with a specific target object because it should encode the information needed to correctly discriminate the target from all the other objects in the scene. In the dialogue of Figure 1, when the model generates a representation for the last turn of the conversation (i.e., “Q: Is it the microwave? A: Yes”), it should encode the fact that “it is an appliance”, “it is not the oven” and “it is the microwave”, allowing the agent to guess the target object correctly.

By playing several guessing games that have a microwave as the target object, the agent should learn a representation of microwave that is expressive enough to correctly discriminate a microwave from all the other objects in a scene. In this setup we are not assuming that the model has a single representation for the concept of microwave; rather the concept of microwave develops from aggregating multimodal information related to microwaves across the situations in which the object is experienced (Barsalou, 2017). In the context of CompGuessWhat?!, every successful dialogue involving a microwave as the target object will be considered as an experience.

We are interested in understanding whether the dialogue state representation generated by a neural model for the last turn of the dialogue can encode the attributes of the target object specified during the dialogue. To do so, we define four attribute prediction tasks. For every target object we predict the corresponding vector composed of: 1) abstract attributes only (A); 2) situated attributes only (S),

---

\[^3\]In the test dataset multiple conversations are associated with the same (image, target object) pair. Therefore, we want the pair (image, target object) to be considered only once in the accuracy evaluation.
3) the union of abstract and situated attributes (AS), and 4) location attributes (L) such as center, top, bottom, right and left. After training the model on the original GuessWhat?! dataset, we can generate dialogue representations corresponding to all the CompGuessWhat?! successful games. Then, we can train a diagnostic classifier that predicts the attributes associated with a given object category using the dialogue hidden representation generated for a given game as features. We hypothesise that a model that has learned grounded representations that are expressive enough to correctly guess the target object should retain the relevant features to predict its attributes.

We treat the attribute-prediction problem as a multi-label classification task. We implement our diagnostic classifier Φ as a linear transformation parameterised by a weight matrix $\mathbb{R}^{d_d \times d_o}$ (where $d_d$ is the dialogue hidden state size and $d_o$ is the number of attributes to be predicted) followed by a sigmoid activation function. We use a sigmoid activation function because it models a Bernoulli distribution. The diagnostic classifier outputs $d_o$ logits where each of them models the probability $P(y_k = 1|d)$ (where $d$ is dialogue state representation), one for each attribute $y_k$ to be predicted. To mitigate a possible class-imbalance problem, we apply a filtering strategy to remove underrepresented attributes from our attribute set, which is a similar technique used to deal with out-of-vocabulary words. We also decided to avoid using class-weighting so that we could evaluate the power of the learned representations with simple linear classifiers as done in previous work using probing classifiers (Belinkov and Glass, 2019). Please refer to Appendix A.3 for details about the procedure to derive the reference set of attributes.

We use the CompGuessWhat?! dataset split as the reference for our training and evaluation setup: we train the diagnostic classifiers on CompGuessWhat?! gold training dialogues and evaluate their performance on the test dialogues using the validation set dialogues for early stopping. We consider Precision, Recall, and F1-measure for multi-label classification (Sorower, 2010) (computed as macro-average) and evaluate them with 0.5 as the threshold value for the sigmoid activation function (selected after considering the models performance using threshold values of 0.75 and 0.9). We report additional details in Appendix A.3.

4.3 Zero-shot Evaluation

Assuming that the model has learned to compose concepts during the turns of the dialogue, we hypothesise that it should also be able to use these representations to play games involving target objects that belong to categories that have never been seen before. For example, humans can discriminate between a dolphin and a dog even though they might not know what it is called. The measure presented in this section has the potential to demonstrate whether current models lack the ability to systematically generalise to new instances that are composed of attributes learned during training.

In order to assess the true generalisation power of the trained agents, we define a zero-shot learning scenario based on the nocaps dataset images (Agrawal et al., 2018). The nocaps dataset is composed of 3 evaluation splits: 1) in-domain: annotated objects belong to MSCOCO categories only; 2) near-domain: contains a mixture of MSCOCO and OpenImages objects; 3) out-of-domain: contains only OpenImages object categories. Since the number of categories in the original GuessWhat?! dataset (80) is lower than the number of categories in the Open Images dataset (660) – contained in nocaps – there are many categories that are never seen during training. Therefore, we can create zero-shot learning games by considering a target object for the game whose category has never been seen during training. We define an automatic procedure to generate the set of reference games for the zero-shot learning setup using the nocaps images. We split the nocaps images into near-domain or out-of-domain. An image is considered near-domain if it contains at least one object whose category belongs to MSCOCO. In contrast, we consider the image out-of-domain if it does not contain any MSCOCO category objects. This procedure generates a dataset of 19, 179 near-domain reference games and 18, 672 out-of-domain reference games. More details about the automatic procedure as well as the resulting reference set of games can be found in Appendix A.4. As a last step of our evaluation framework, we evaluate the performance of the state-of-the-art models in the zero-shot gameplay setup. For this task, the trained models need to interact with each other and generate dialogues given the pair (image, target object). As an evaluation metric for this task, we consider gameplay guesser accuracy for the near-domain (ND-Acc) and out-of-domain (OD-Acc) reference games.
Table 1: Results for state-of-the-art models on the CompGuessWhat?! suite of evaluation tasks. We assess model quality in terms of gameplay accuracy, the attribute prediction quality, measured in terms of F1 for the abstract (A-F1), situated (S-F1), abstract+situated (AS-F1) and location (L-F1) prediction scenario, as well as zero-shot learning gameplay. The final score GroLLA is a macro-average of the individual scores. We use the models GloVe, ResNet and GDSE-**-text only as a baseline for the attribute prediction tasks.

<table>
<thead>
<tr>
<th></th>
<th>Gameplay Accuracy</th>
<th>Attribute Prediction</th>
<th>Zero-shot Gameplay</th>
<th>GroLLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A-F1</td>
<td>S-F1</td>
<td>AS-F1</td>
</tr>
<tr>
<td>Random</td>
<td>15.81%</td>
<td>15.1</td>
<td>0.1</td>
<td>7.8</td>
</tr>
<tr>
<td>GloVe</td>
<td>-</td>
<td>34.6</td>
<td>29.7</td>
<td>36.4</td>
</tr>
<tr>
<td>ResNet</td>
<td>-</td>
<td>24.5</td>
<td>31.7</td>
<td>27.9</td>
</tr>
<tr>
<td>GDSE-SL-text</td>
<td>-</td>
<td>57.0</td>
<td>45.3</td>
<td>57.5</td>
</tr>
<tr>
<td>GDSE-CL-text</td>
<td>-</td>
<td>56.9</td>
<td>45.0</td>
<td>57.3</td>
</tr>
<tr>
<td>DeVries-SL</td>
<td>41.5%</td>
<td>46.8</td>
<td>39.1</td>
<td>48.5</td>
</tr>
<tr>
<td>DeVries-RL</td>
<td>53.5%</td>
<td>45.2</td>
<td>38.9</td>
<td>47.2</td>
</tr>
<tr>
<td>GDSE-SL</td>
<td>49.1%</td>
<td>59.9</td>
<td>47.6</td>
<td>60.1</td>
</tr>
<tr>
<td>GDSE-CL</td>
<td>59.8%</td>
<td>59.5</td>
<td>47.6</td>
<td>59.8</td>
</tr>
</tbody>
</table>

5 Results: Model Evaluation using CompGuessWhat?!

Guesser accuracy We evaluate the GDSE and DeVries models in gameplay mode using the set of reference games provided in CompGuessWhat?!. As shown in Table 1, the results are in line with the performance of the models on the original GuessWhat?! dataset (de Vries et al., 2017; Shekhar et al., 2019) confirming that our filtering strategy did not affect the complexity of the task.

Attribute Prediction We use the CompGuessWhat?! benchmark to compare several dialogue state representations:

- DeVries-SL: the representation learned by the Questioner model presented in (de Vries et al., 2017) that generates the question tokens conditioned on the image features and is trained using Supervised Learning (SL).
- DeVries-RL: the representations learned by the Questioner model presented in (de Vries et al., 2017), fine-tuned using the Reinforcement Learning procedure proposed in (Strub et al., 2017).
- GDSE-SL: the grounded dialogue state learned by a seq2seq model trained using the multi-task Learning procedure in (Shekhar et al., 2019).
- GDSE-CL: the grounded dialogue state learned by the Questioner model used in GDSE-SL, fine-tuned with the Collaborative Learning procedure presented in (Shekhar et al., 2019).
- GDSE-SL-text: the learned LSTM (Hochreiter and Schmidhuber, 1997) dialogue encoder of the GDSE-SL model.
- GDSE-CL-text: the learned dialogue encoder of the GDSE-CL model.

In order to control for possible bias in our task, we consider unimodal (Thomason et al., 2019a) as well as random attribute predictors:
- GloVe: a dialogue is represented as the average of the GloVe embeddings associated with each word (Pennington et al., 2014).
- ResNet: uses the latent representation of the reference scene generated by a ResNet152 as proposed in Shekhar et al. (2019).
- Random: samples $d_a$ scores from $U(0, 1)$ where samples are independent from each other. We incorporate this baseline as a lower bound performance on the attribute prediction task.

With the AP task, we try to answer the following question: “Do the representations associated with the target object encoding provide useful information that can be exploited to predict the object attributes correctly?” We assume that, due to the nature of the CompGuessWhat?! games, the final dialogue state representation should encode relevant features of the target object. So, a high gameplay accuracy should correlate with a high AP score. Table 1 summarises the results of the attribute prediction task evaluated on the CompGuessWhat?!

4We could use the dialogue encoder of the GDSE models only due to their modular architecture. It was not possible to properly separate the dialogue encoder from the visual representation in the DeVries models.
test games. As the average best model performance was only 44.27, far from ceiling, our hypothesis is only partially supported. In particular, the models having the highest guesser accuracy, GDSE-CL and GDSE-SL, seem to learn better representations than unimodal baselines GloVe and ResNet, confirming the importance of multi-modal training for this task. There is also a gap in performance between the GDSE and DeVries models. This might be related to the multi-task learning strategy used by GDSE models that favours the emergence of more expressive representations than the ones learned by DeVries models which are trained in isolation. By comparing the enhanced versions GDSE-CL and DeVries-RL with the less sophisticated ones, GDSE-SL and DeVries-SL, respectively, we observe that, despite their higher guesser accuracy, these models do not have any advantage in terms of the AP task. We believe that this is because the Collaborative training strategy (for GDSE-CL) and Reinforcement Learning (for DeVries-RL) are optimising end-goal performance while sacrificing the expressiveness of the representations. Finding a way to encode task-specific representations and generalise them to learn abstract representations becomes an important research direction to improve on this task.

As an additional ablation, we compared the representations learned by the LSTM module used by GDSE to encode the dialogue (GDSE-\texttt{+text}) with their grounded dialogue state counterpart. Differences in terms of F1 are minimal, confirming that the heavy lifting is done by the textual representations and it is not clear how well the grounded dialogue state retains the visual information. Another confirmation of this issue is provided by the results in terms of location attributes prediction. Performance in this task for all the models is around 40% meaning that both VGGNet and ResNet features (used for DeVries and GDSE, respectively) are not able to recover fine-grained object information. This result sheds light on the ability of these models to ground the textual data in perceptual information of the reference scene. We believe that models should be able to co-ground one modality with the other and, as a result, learn more expressive grounded representations.

Zero-shot Evaluation Results are summarised in Table 1; the most striking observation is that all models struggle with this dataset (guesser accuracy is barely above 40), although arguably humans would be able to solve the task despite their unfamiliarity with a specific object. Indeed, in this zero-shot scenario, reusing previously learned attributes that are shared among the objects or leveraging mutual exclusivity (Markman and Wachtel, 1988) would result in a successful gameplay.

Even the most accurate model in the CompGuess-What?! guesser evaluation performs poorly in this zero-shot setup (see Figure 3 for an example). We attribute this drop in performance to the way that these models represent objects. In particular, they all rely on category embeddings, i.e., latent representations associated to specific object categories (refer to (Shekhar et al., 2019; de Vries et al., 2017) for more details). In the case of ZS evaluation, when an object is unknown, its category embedding is also not available. This is true for both DeVries and GDSE models; it seems that GDSE models suffer more than DeVries models possibly due to overfitting. On the other hand, we aim to learn object representations which are not associated with manually-provided categories but are obtained by playing the game and that encode both abstract and situated attributes.

Once again, we find that models optimised using Reinforcement Learning seem to learn a better game strategy that results in higher performance on both near-domain and out-of-domain games. To better understand the quality of the generated dialogues, we classify each type of question according to a pre-defined set of types based on (Shekhar et al., 2019) (please refer to Appendix A.5 for a detailed description and a detailed summary of the evaluation results). We noticed that the DeVries models generate dialogues with 70% of their turns comprising “location” questions (e.g., “is it the person on the right?”) compared to 20% for GDSE models. We argue that to tackle zero-shot scenes, a model should instead learn features useful to discriminate the target object without relying on locations. Of course, in some reference scenes, location questions are still useful attributes used by humans when playing the game. In addition, asking location questions is an effective strategy because the Oracle has access to positional information that can be used to provide reliable answers but does not have any category embeddings for the target object.

6 Related Work

Text-only Multi-task Evaluation In recent years progress in NLP has been driven by multi-
task evaluation datasets proposed to mitigate the biases of task-specific datasets (Wang et al., 2018; McCann et al., 2018; Wang et al., 2019). Despite their multi-task nature, these datasets focus on text-only data making the resulting models unable to learn meaning representations which are grounded in perceptual experience (Andrews et al., 2009; Riordan and Jones, 2011). Another downside is that these benchmarks focus only on end-goal metrics, i.e., are not informative on what the model has learned. Going beyond the end-goal metric is fundamental for designing models that are more generalisable and interpretable. By introducing the attribute prediction task in our framework, we assess whether the learned representations are expressive enough to predict the attributes of relevant objects in the scene. Also, we propose a zero-shot evaluation where the model has to generate predictions for examples that have never been seen during training, thus providing a way to understand the generalisation power of the learned representations.

**Grounded Language Learning Evaluation**
Several grounded language learning tasks have been proposed in the literature that can be divided into discriminative (Shekhar et al., 2017; Hudson and Manning, 2019; Suhr et al., 2019) and generative grounded language learning tasks (Xu et al., 2015; Agrawal et al., 2018). Recent works proposed models trained in a multi-task fashion by exploiting several language/vision tasks. The dodecaDialogue task (Shuster et al., 2019) proposes twelve dialogue tasks, among which there are two language/vision tasks in which the agent has to generate a response for a given context. Other works try to exploit multi-task learning to improve on single-task model performance in discriminative tasks (Pramanik et al., 2019; Lu et al., 2019). Unfortunately, implementing multi-task learning using different datasets results is cumbersome (Subramanian et al., 2018). We propose an evaluation framework that can be applied in the context of a single task and dataset (e.g. GuessWhat?! ) that allows to understand the extent to which the model can learn useful representations for the task at hand.

Inspecting the learned representations is important because, due to biases in the datasets, models might learn spurious correlations between input and output rather than actual grounding capabilities. For instance, in Visual Question Answering, questions starting with “What colour are” have “white” as a correct answer 23% of the time; models learn to memorise this sort of association rather than using the visual information (Zhang et al., 2016; Agrawal et al., 2016). This issue calls for a model evaluation aimed at inspecting the model representations as well as how these representations are used. The GQA (Hudson and Manning, 2019) dataset goes in this direction. It presents a Visual Question Answering dataset where images are supported by rich semantic annotations in the form of scene graphs. The GQA task requires the model to select an answer among a set of candidates. However, we advocate the importance of tasks that involve both Natural Language Understanding (NLU) and Natural Language Generation (NLG) skills in a curriculum for grounded language learning. There are significant differences concerning the proposed auxiliary tasks as well. First of all, GQA’s tasks are specifically designed around the VQA tasks to make sure that the model is consistent and plausible. It does not however tell us what the model’s learned representations are encoding.

We propose the AP task as a diagnostic task
aimed at better understanding the learned neural representations (Belinkov and Glass, 2017; Conneau et al., 2018; Peters et al., 2018; Tenney et al., 2019). In addition, going beyond simple object classification is considered beneficial for vision systems (Farhadi et al., 2009) because it allows generalisation across object categories, not just across instances within a category. However, we believe that to truly assess the generalisation ability of a model, object attributes have to be used for the downstream task, which is not necessarily needed in object classification tasks. With the ZS evaluation, we investigate the ability of the models to exploit more fine-grained visual attributes which is important for models able to learn from few examples and easily transfer to new domains.

**Compositionality Evaluation** Andreas (2019) presents a method to estimate the degree of compositionality of neural representations by using an oracle compositional model aware of the compositional structure (i.e., a derivation) of the input data. Building a reference oracle is easy for synthetic scenes (as in Andreas (2019)) but is a significant challenge for real-world scenes. Previous work has studied compositionality in real-world scenes for visual concept composition (Misra et al., 2017) and image captioning (Nikolaus et al., 2019). In our benchmark CompGuessWhat?! we use real-world scenes from the MSCOCO (Lin et al., 2014) and OpenImages (Kuznetsova et al., 2018) datasets. Our AP task is related to measuring compositionality. It relies on image annotations in the form of intensional and extensional attributes as a reference structure for the objects in the scene.

### 7 Conclusions & Future Work

We proposed CompGuessWhat?! as an implementation of GroLLA, a multi-task evaluation framework for Grounded Language Learning with Attributes. We found that the best performing model achieves a GroLLA score of 50.06%; notably this model’s out-of-domain accuracy is under 30%, as compared to the human performance on the original GuessWhat?! dataset of 90.2% (de Vries et al., 2017). Clearly, even models with high in-domain gameplay success rates still have difficulty generalising to new scenarios. In the following, we discuss insights gained from the evaluation and new research directions for this task.

The attribute prediction task shows that model representations are not able to accurately recover attribute representations. We argue that this result calls for new approaches to exploiting and representing textual and visual data. We believe that models should be equipped with a co-grounding operator that fuses the textual and visual modalities. For instance, in the context of CompGuessWhat?! it would be used to learn a representation for the current turn that is influenced by both the language and visual modality. CompGuessWhat?! requires models to learn to combine the co-grounded information provided for every turn. Therefore, we propose that CompGuessWhat?! represents a benchmark dataset for evaluating the design of such an attribute compositionality operator that would be a possible implementation of compositionality à la Barsalou (2017).

In this work, we have shown how our multi-task evaluation framework can be applied to GuessWhat?! However, the same framework could be applied to other multi-modal tasks. For example, in image captioning, the goal-oriented evaluation would be the textual similarity metrics (e.g. BLEU); the attribute-prediction task would use the decoder representation to predict the attributes of the objects in the image (Elliott and Kádár, 2017, e.g.); and the zero-shot setting could leverage the nocaps dataset (Agrawal et al., 2018). Likewise, in the Vision-and-Dialog navigation task (Thomason et al., 2019b), the goal-oriented evaluation is the navigation task; attribute prediction is based on predicting the attributes of the hidden object when the agent decides it is in the correct room, and the zero-shot setting could evaluate model performance on novel combinations of rooms and object types.

Finally, from the evaluation presented here, it emerges that these models learn task-specific representations that do not generalise to unseen object categories. We hope that GroLLA and the CompGuessWhat?! data will encourage the implementation of learning mechanisms that fuse task-specific representations with more abstract representations to encode attributes in a more compositional manner. In addition, we will use the CompGuessWhat?! image annotations to design a visual grounding evaluation to assess the ability of the model to attend to the correct objects during the turns of the dialogue.

### Acknowledgements

We thank Arash Eshghi and Yonatan Bisk for fruitful discussions in the early stages of the project.
References


A Appendices

A.1 CompGuessWhat?! Dataset

We extend the GuessWhat?! dataset (de Vries et al., 2017) to promote the study of compositional grounded language representations. The original GuessWhat?! dataset has been enhanced by including a semantic layer on top of the purely perceptual one (i.e., images). In particular, we enrich the VisualGenome (Krishna et al., 2017) scene graphs associated with the GuessWhat?! images with several attributes coming from resources such as VISA (Sillerer and Lapata, 2012) and ImSitu (Yatskar et al., 2016). As shown in Table 2 not all the GuessWhat?! images are included in VisualGenome. We were able to reuse 40.79% of the original GuessWhat?! dialogues for a total of 65,700 dialogues as summarised in Table 3.

<table>
<thead>
<tr>
<th>Split</th>
<th>GuessWhat?! images</th>
<th>Mapped images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>46794</td>
<td>19117</td>
</tr>
<tr>
<td>Validation</td>
<td>9844</td>
<td>4049</td>
</tr>
<tr>
<td>Test</td>
<td>9899</td>
<td>3989</td>
</tr>
</tbody>
</table>

Table 2: Statistics of the mapping between GuessWhat?! images and VisualGenome images.

A.2 VisualGenome object mapping

VisualGenome images are not exactly the same in terms of shape and content as the ones in MSCOCO. This is due to the presence of possible visual elements (i.e., banners) that are in the VisualGenome version of the image and are not in the MSCOCO one. This complicates the object mapping procedure used to link together abstract attributes and attributes coming from VisualGenome. As a first step, the procedure finds the largest VisualGenome bounding box with an IoU greater than 0.5. If there is not one, it looks for the largest VisualGenome bounding box with an IoU which is not close to 0 (with a tolerance of 0.05) and whose category is similar to the one of the reference MSCOCO one (where ‘similar’ is measured according to the Jaccard index between the corresponding category tokens). Whenever the MSCOCO object bounding box cannot be mapped to one of the VisualGenome bounding boxes, we assume that we do not have access to the situated attributes and we use the abstract attributes associated to its MSCOCO category only.

A.3 Diagnostic Classifiers for Attribute Prediction

For the attribute prediction task we apply a filtering procedure on the attribute set that will be used for training. In particular, we ignore all the attributes that belong to the abstract attribute category whose frequency is less than 100 (resulting in a set of attributes equal to 1997) and we ignore all the situated attributes whose frequency is less than 2 (resulting in a set of attributes equal to 4085).

For the attribute-prediction task we define a probing classifier \( \Phi \) as a linear transformation parameterised by a weight matrix \( \mathbb{R}^{d_d \times d_a} \) (where \( d_d \) is the dialogue hidden state size and \( d_a \) is the number of attributes to be predicted) followed by a sigmoid activation function. The number of input dimensions \( d_d \) depends on the model hidden state representations. We report in Table 4 the corresponding hidden state sizes for all the evaluated models. The output size \( d_a \) depends on the attribute set that we intend to consider. When situated attributes are considered \( d_a = 6082, d_a = 1997 \) for abstract attributes, \( d_a = 5 \) for location attributes and \( d_a = 4085 \) for situated-only attributes.

We consider the CompGuessWhat?! splits as reference for our experimental evaluation. We generate an hidden state for every successful dialogue and we use the classifier \( \Phi \) to predict the target object attributes. We train the classifier \( \Phi \) by minimising the binary cross-entropy loss computed between the model prediction and the reference set of attributes. We use ADAM (Kingma and Ba, 2014) as optimiser for our training procedure. To prevent overfitting, we perform early stopping on the validation set using the multi-label F1-measure (with threshold 0.75) as reference metric and we apply a learning rate scheduler to gradually reduce the learning rate. The model training has been implemented using AllenNLP (Gardner et al., 2018). We report the full set of metrics evaluated for this task in Table 5.

For the GDSE models we used a modified version of the code provided by the author via personal correspondence. On the other hand, for the DeVries model, we use the pretrained models and code that is available on the official webpage.

\footnote{https://github.com/GuessWhatGame/}
Table 3: Comparison between the original GuessWhat?! dataset and CompGuessWhat?! dataset. We report the percentage of dialogues that we retain after the filtering procedure based on the VisualGenome images.

<table>
<thead>
<tr>
<th>Split</th>
<th># GuessWhat?! dialogues</th>
<th># CompGuessWhat?! dialogues</th>
<th>Vocab. size</th>
<th>Avg. dialogue length</th>
<th>Successful dialogues</th>
<th>Failed dialogues</th>
<th>Incomplete dialogues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>113221</td>
<td>46277</td>
<td>7090</td>
<td>5.128</td>
<td>84.06% (38901)</td>
<td>10.35% (4790)</td>
<td>5.59% (2586)</td>
</tr>
<tr>
<td>Valid</td>
<td>23739</td>
<td>9716</td>
<td>3605</td>
<td>5.106</td>
<td>83.97% (8159)</td>
<td>11.03% (1069)</td>
<td>5.03% (488)</td>
</tr>
<tr>
<td>Test</td>
<td>23785</td>
<td>9619</td>
<td>3552</td>
<td>5.146</td>
<td>84.10% (8090)</td>
<td>10.74% (1034)</td>
<td>5.14% (495)</td>
</tr>
</tbody>
</table>

Table 4: Summary of hidden state sizes for all the models considered in the attribute prediction evaluation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hidden size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeVries-SL</td>
<td>512</td>
</tr>
<tr>
<td>DeVries-RL</td>
<td>512</td>
</tr>
<tr>
<td>GDSE-SL</td>
<td>512</td>
</tr>
<tr>
<td>GDSE-CL</td>
<td>512</td>
</tr>
<tr>
<td>GDSE-SL-text</td>
<td>1024</td>
</tr>
<tr>
<td>GDSE-CL-text</td>
<td>1024</td>
</tr>
<tr>
<td>GloVe</td>
<td>300</td>
</tr>
<tr>
<td>ResNet</td>
<td>2048</td>
</tr>
</tbody>
</table>

A.4 Zero-shot Evaluation Reference Games Generation

We define an automatic procedure to generate the set of reference games for the zero-shot learning setup. Specifically, for all the images in nocaps validation and test sets we first extract all the bounding boxes that satisfy the following conditions: 1) bounding box area should be greater than 500 pixels; 2) bounding box region should not be occluded; 3) bounding box region should not be truncated; 4) bounding box should not be associated with human body parts. An additional inclusion condition for the image is that the number of valid bounding boxes should be between 3 and 20. This ‘sanity check’ step is inspired by the procedure adopted in the original GuessWhat?! dataset (de Vries et al., 2017) and is used in order to guarantee that the gameplay reference images are not too crowded or composed of really small objects. Finally, we split the valid images in near-domain or out-of-domain. An image is considered near-domain if it contains at least an object whose category belongs to MSCOCO; we consider the image out-of-domain if it does not contain any MSCOCO category.

All the valid images resulting from the sanity check step can be considered as reference scene for the game. In order to define a fair comparison between all the agents, we define a reference set of games by sampling a fixed number of target objects for every image (e.g., 5 objects). In order to make sure that the sampling procedure is not biased by the frequency of the classes in the dataset, we sample an object according to the inverse of its category frequency in the dataset. As a result of this procedure, as shown in Table 6, we generated a dataset of 19,179 near-domain reference games and 18,672 out-of-domain reference games. In Figure 4 and 5 show the object category distribution in the near-domain and out-of-domain reference games, respectively.
Figure 4: Object category distribution in the near-domain reference set of games.

Figure 5: Object category distribution in the out-of-domain reference set of games.
### Table 5: Full set of attribute prediction task metrics. We evaluate F1, Precision and Recall for all the tasks. All the metrics are computed as macro-average.

<table>
<thead>
<tr>
<th>Models</th>
<th>Abstract F1</th>
<th>Abstract Recall</th>
<th>Situated-only F1</th>
<th>Situated-only Recall</th>
<th>Abstract-situated F1</th>
<th>Abstract-situated Recall</th>
<th>Location F1</th>
<th>Location Precision</th>
<th>Location Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeVries-SL</td>
<td>46.8</td>
<td>53.4</td>
<td>39.1</td>
<td>51.2</td>
<td>48.5</td>
<td>57.8</td>
<td>42.7</td>
<td>42.8</td>
<td>42.9</td>
</tr>
<tr>
<td>DeVries-RL</td>
<td>45.2</td>
<td>52.5</td>
<td>38.9</td>
<td>51</td>
<td>47.2</td>
<td>57.4</td>
<td>43.5</td>
<td>43.6</td>
<td>43.6</td>
</tr>
<tr>
<td>GDSE-SL</td>
<td>59.9</td>
<td>64.1</td>
<td>47.6</td>
<td>60.1</td>
<td>63.8</td>
<td>65.9</td>
<td>48.3</td>
<td>48.6</td>
<td>48.6</td>
</tr>
<tr>
<td>GDSE-CL</td>
<td>59.5</td>
<td>63.6</td>
<td>47.6</td>
<td>59.8</td>
<td>63.3</td>
<td>65.6</td>
<td>48.1</td>
<td>48.1</td>
<td>48.6</td>
</tr>
<tr>
<td>GDSE-SL-text</td>
<td>57</td>
<td>61.5</td>
<td>45.3</td>
<td>57.5</td>
<td>60.6</td>
<td>60.6</td>
<td>46</td>
<td>46.1</td>
<td>46.4</td>
</tr>
<tr>
<td>GDSE-CL-text</td>
<td>56.9</td>
<td>61.4</td>
<td>45</td>
<td>57.3</td>
<td>60.5</td>
<td>60.5</td>
<td>45</td>
<td>45</td>
<td>45.4</td>
</tr>
<tr>
<td>GloVe</td>
<td>34.6</td>
<td>45.9</td>
<td>29.7</td>
<td>36.4</td>
<td>37.4</td>
<td>52.9</td>
<td>33.6</td>
<td>33.6</td>
<td>33.7</td>
</tr>
<tr>
<td>ResNet</td>
<td>24.5</td>
<td>37.9</td>
<td>31.7</td>
<td>27.9</td>
<td>30.3</td>
<td>47.1</td>
<td>43.4</td>
<td>43.5</td>
<td>43.6</td>
</tr>
<tr>
<td>Random</td>
<td>15.1</td>
<td>16.3</td>
<td>0.1</td>
<td>7.8</td>
<td>50.3</td>
<td>5.4</td>
<td>27.5</td>
<td>49.7</td>
<td>20.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># images</th>
<th># games</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near-domain validation</td>
<td>1208</td>
</tr>
<tr>
<td>Out-of-domain validation</td>
<td>1306</td>
</tr>
<tr>
<td>Near-domain test</td>
<td>3097</td>
</tr>
<tr>
<td>Out-of-domain test</td>
<td>3212</td>
</tr>
</tbody>
</table>

Table 6: Statistics for the CompGuessWhat?! zero-shot scenario. We provide near-domain and out-of-domain splits using specific nocaps images as reference scenes.

### A.5 Generated Dialogue Evaluation

In order to evaluate to provide a more fine-grained evaluation of the generated dialogues, we adapt the quality evaluation script presented by (Shekhar et al., 2019) and extend it with additional metrics. First of all, it relies on a rule-based question classifier that classifies a given question in one of seven classes: 1) super-category (e.g., “person”, “utensil”, etc.), 2) object (e.g., “car”, “oven”, etc.), 3) “color”, 4) “size”, 5) “texture”, 6) “shape” and “location”. The question classifier is useful to evaluate the dialogue strategy learned by the models. In particular, we look at two types of turn transitions: 1) super-category $\rightarrow$ object/attr, it measures how many times a question with an affirmative answer from the Oracle related to a super-category is followed by either an object or attribute question (where “attribute” represents the set \{color, size, texture, shape and location\}); 2) object $\rightarrow$ attr, it measures how many times a question with an affirmative answer from the Oracle related to an object is followed by either an object or attribute question. We compute the *lexical diversity* as the type/token ratio among all games, *question diversity* and the percentage of games with repeated questions. We also evaluate the percentage of dialogue turns involving location questions. Table 7 and 8 show the results of these analysis for the models DeVries and GDSE analysed in this paper.
<table>
<thead>
<tr>
<th>Model</th>
<th>Lexical diversity</th>
<th>Question diversity</th>
<th>% games repeated questions</th>
<th>Super-cat -&gt; obj/attr</th>
<th>Object -&gt; attribute</th>
<th>% turns location questions</th>
<th>Vocab. size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeVries-SL</td>
<td>0.76</td>
<td>44.64</td>
<td>12.54%</td>
<td>97.33%</td>
<td>73%</td>
<td>29.34%</td>
<td>2668</td>
<td>31.33%</td>
</tr>
<tr>
<td>DeVries-RL</td>
<td>0.13</td>
<td>1.77</td>
<td>99.48%</td>
<td>96.43%</td>
<td>98.63%</td>
<td>78.07%</td>
<td>702</td>
<td>43.92%</td>
</tr>
<tr>
<td>GDSE-SL</td>
<td>0.13</td>
<td>6.10</td>
<td>92.38%</td>
<td>95.60%</td>
<td>52.35%</td>
<td>15.74%</td>
<td>862</td>
<td>29.78%</td>
</tr>
<tr>
<td>GDSE-CL</td>
<td>0.17</td>
<td>13.74</td>
<td>66.76%</td>
<td>99.48%</td>
<td>67.25%</td>
<td>31.23%</td>
<td>1260</td>
<td>43.42%</td>
</tr>
</tbody>
</table>

Table 7: Gameplay quality analysis on Near-domain zero-shot reference games.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lexical diversity</th>
<th>Question diversity</th>
<th>% games repeated questions</th>
<th>Super-cat -&gt; obj/attr</th>
<th>Object -&gt; attribute</th>
<th>% turns location questions</th>
<th>Vocab. size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeVries-SL</td>
<td>0.83</td>
<td>45.86</td>
<td>11.58</td>
<td>97.87%</td>
<td>76.50%</td>
<td>29.64%</td>
<td>2604</td>
<td>28.37%</td>
</tr>
<tr>
<td>DeVries-RL</td>
<td>0.24</td>
<td>2.96</td>
<td>98.49%</td>
<td>91.83%</td>
<td>98.58%</td>
<td>75.84%</td>
<td>1275</td>
<td>38.73%</td>
</tr>
<tr>
<td>GDSE-SL</td>
<td>0.09</td>
<td>1.31</td>
<td>97.19%</td>
<td>100%</td>
<td>67.45%</td>
<td>7.90%</td>
<td>519</td>
<td>22.32%</td>
</tr>
<tr>
<td>GDSE-CL</td>
<td>0.14</td>
<td>7.86</td>
<td>66.32%</td>
<td>100%</td>
<td>71.14%</td>
<td>26.03%</td>
<td>1002</td>
<td>29.83%</td>
</tr>
</tbody>
</table>

Table 8: Gameplay quality analysis on Out-of-domain zero-shot reference games.